Abstract of the Disclosure

A multi-lingual indexing and search system performs tokenization and stemming in a manner which is independent of whether index entries and search terms appear as words in a dictionary. During the tokenization phase of the process, a string of text is separated into individual word tokens, and predetermined types of tokens are eliminated from further processing. The stemming phase of the process reduces words to grammatical stems by removing known word-endings associated with the various languages to be supported. Known word endings are removed from the word tokens without any effort to guarantee that the remaining stem is contained in a dictionary. In a preferred implementation, the stemming process is only applied to nouns.